Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas

Mager, Manuel ; Oncevay, Arturo ; Ebrahimi, Abteen ; Ortega, John ; Rios, Annette ; Fan, Angela ; Gutierrez-Vasques, Ximena ; Chiruzzo, Luis ; Giménez-Lugo, Gustavo ; Ramos, Ricardo ; Meza Ruiz, Ivan Vladimir ; Coto-Solano, Rolando ; Palmer, Alexis ; Mager-Hois, Elisabeth ; Chaudhary, Vishrav ; Neubig, Graham ; Vu, Ngoc Thang ; Kann, Katharina

Abstract: This paper presents the results of the 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. The shared task featured two independent tracks, and participants submitted machine translation systems for up to 10 indigenous languages. Overall, 8 teams participated with a total of 214 submissions. We provided training sets consisting of data collected from various sources, as well as manually translated sentences for the development and test sets. An official baseline trained on this data was also provided. Team submissions featured a variety of architectures, including both statistical and neural models, and for the majority of languages, many teams were able to considerably improve over the baseline. The best performing systems achieved 12.97 ChrF higher than baseline, when averaged across languages.

Originally published at:
Mager, Manuel; Oncevay, Arturo; Ebrahimi, Abteen; Ortega, John; Rios, Annette; Fan, Angela; Gutierrez-Vasques, Ximena; Chiruzzo, Luis; Giménez-Lugo, Gustavo; Ramos, Ricardo; Meza Ruiz, Ivan Vladimir; Coto-Solano, Rolando; Palmer, Alexis; Mager-Hois, Elisabeth; Chaudhary, Vishrav; Neubig, Graham; Vu, Ngoc Thang; Kann, Katharina (2021). Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Online, 11 June 2021. Association for Computational Linguistics, 202-217.

NAACL-HLT 2021

Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)

Proceedings of the First Workshop

June 11, 2021

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-954085-44-2

Preface

This area is in all probability unmatched, anywhere in the world, in its linguistic multiplicity and diversity. A couple of thousand languages and dialects, at present divided into 17 large families and 38 small ones, with several hundred unclassified single languages, are on record. In one small portion of the area, in Mexico just north of the Isthmus of Tehuantepec, one finds a diversity of linguistic type hard to match on an entire continent in the Old World.
—McQuown (1955)

Workshop Organizers:

Manuel Mager
Arturo Oncevay
Annette Rios
Ivan Vladimir Meza Ruiz
Alexis Palmer
Graham Neubig
Katharina Kann

Shared Task Organizers:

Manuel Mager
Arturo Oncevay
Abteen Ebrahimi
John Ortega
Annette Rios
Angela Fan
Ximena Gutierrez-Vasques
Luis Chiruzzo
Gustavo A. Giménez-Lugo
Ricardo Ramos
Ivan Vladimir Meza Ruiz
Rolando Coto-Solano
Alexis Palmer
Elisabeth Mager
Vishrav Chaudhary
Graham Neubig
Ngoc Thang Vu
Katharina Kann

Program Committee:

Abhilasha Ravichander
Abteen Ebrahimi
Adam Wiemerslage
Alexandra Birch
Alfonso Medina-Urrea
Annette Rios
Antonios Anastasopoulos
Arturo Oncevay
Barry Haddow
Candace Ross
Cynthia Montaño
Daan van Esch
Ekaterina Vylomova
Emily M. Bender
Emily Prud'hommeaux
Eneko Agirre
Fernando Alva-Manchego
Francis Tyers
Gerardo Sierra Martínez
Ivan Vulić

John Miller
Judith Klavans
Luke Gessler
Manuel Mager
Marco Antonio Sobrevilla Cabezudo
Nickil Maveli
Pavel Denisov
Rico Sennrich
Robert Pugh
Roberto Zariquiey
Ronald Cardenas
Sarah Moeller
Shruti Rijhwani
Taraka Rama
William Abbott Lane
Ximena Gutierrez-Vasques
Yoshinari Fujinuma
Zoey Liu

Table of Contents

Workshop Program

June 11, 2021

qxoRef 1.0: A coreference corpus and mention-pair baseline for coreference resolution in Conchucos Quechua
Elizabeth Pankratz

A corpus of K'iche' annotated for morphosyntactic structure
Francis Tyers and Robert Henderson

Investigating variation in written forms of Nahuatl using character-based language models
Robert Pugh and Francis Tyers

Apurinã Universal Dependencies Treebank
Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen and Niko Partanen

Automatic Interlinear Glossing for Otomi language
Diego Barriga Martínez, Victor Mijangos and Ximena Gutierrez-Vasques

A survey of part-of-speech tagging approaches applied to K'iche'
Francis Tyers and Nick Howell

Highland Puebla Nahuatl Speech Translation Corpus for Endangered Language Documentation
Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan and Shinji Watanabe

End-to-End Automatic Speech Recognition: Its Impact on the Workflowin Documenting Yoloxóchitl Mixtec
Jonathan D. Amith, Jiatong Shi and Rey Castillo García

A finite-state morphological analyser for Paraguayan Guaraní
Anastasia Kuznetsova and Francis Tyers

Morphological Segmentation for Seneca
Zoey Liu, Robert Jimerson and Emily Prud'hommeaux

Representation of Yine [Arawak] Morphology by Finite State Transducer Formalism
Adriano Ingunza Torres, John Miller, Arturo Oncevay and Roberto Zariquiey Biondi

June 11, 2021 (continued)

Leveraging English Word Embeddings for Semi-Automatic Semantic Classification in Nêhiyawêwin (Plains Cree)
Atticus Harrigan and Antti Arppe

Restoring the Sister: Reconstructing a Lexicon from Sister Languages using Neural Machine Translation
Remo Nitschke

Expanding Universal Dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik
Hyunji Park, Lane Schwartz and Francis Tyers

The More Detail, the Better? – Investigating the Effects of Semantic Ontology Specificity on Vector Semantic Classification with a Plains Cree / nêhiyawêwin Dictionary
Daniel Dacanay, Atticus Harrigan, Arok Wolvengrey and Antti Arppe

Experiments on a Guarani Corpus of News and Social Media
Santiago Góngora, Nicolás Giossa and Luis Chiruzzo

Towards a First Automatic Unsupervised Morphological Segmentation for Inuinnaqtun
Ngoc Tan Le and Fatiha Sadat

Toward Creation of Ancash Lexical Resources from OCR
Johanna CORDOVA and Damien Nouvel

Ayuuk-Spanish Neural Machine Translator
Delfino Zacarías Márquez and Ivan Vladimir Meza Ruiz

Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri
Rolando Coto-Solano

Towards a morphological transducer and orthography converter for Western Tlacolula Valley Zapotec
Jonathan Washington, Felipe Lopez and Brook Lillehaugen

Peru is Multilingual, Its Machine Translation Should Be Too?
Arturo Oncevay

Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas
Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu and Katharina Kann

June 11, 2021 (continued)